Access to Freely Available Journal Articles: Gold, Green, and Rogue Open Access Across the Disciplines

Michael Levine-Clark
University of Denver

John McDonald
University of Southern California

Jason Price
SCELC Library Consortium

Michael Levine-Clark, John McDonald, and Jason Price, "Access to Freely Available Journal Articles: Gold, Green, and Rogue Open Access Across the Disciplines" (2016). *Proceedings of the Charleston Library Conference.*
http://dx.doi.org/10.5703/1288284316494

Access to Freely Available Journal Articles: Gold, Green, and Rogue Open Access Across the Disciplines

Michael Levine-Clark, Dean of Libraries, University of Denver

John McDonald, Associate Dean for Collections, University of Southern California

Jason Price, Director of Licensing Operations, SCELC Library Consortium

The following is a transcript of a live presentation at the 2016 Charleston Conference.

Michael Levine-Clark: We're exploring the access and discovery to freely available articles, and we're deliberately looking at not just open access content but anything that is freely available to a user on the web. From a user perspective they might care philosophically whether it is open access versus something that they are getting pirated access to, but the reality is that they may often not even know which type of access it is. So, we're looking at gold open access, green open access, and rogue and pirate open access, stuff that maybe you shouldn't quite have access to.

The library, we know, for many users is not the starting point. A recent ITHAKA report, as well as the New Media Consortium Horizon Report, has talked about this issue that users start very often from Google, from Google Scholar. They don't start from library sources. The ITHAKA report talks about the fact that while discovery services for students are often important, much more often they are starting their searches from other places from the open web. And we've got data that backs that up. This is referral data to a particular publisher (see Figure 1).

The pie chart is the University of Denver, my institution, and this is almost a year's worth of data for a particular publisher, and this is to the licensed content that we have at the University of Denver. Thirty-nine percent of the referrals to our context, to this publisher's content, came from our library discovery services. So, from the discovery service, from the resolver, from the catalog, from databases; so library tools broadly speaking. Sixty-one percent came from other places, right? So, 32% came from Google and Google Scholar together. Twenty-seven percent were not sure where it came from; there is no clear originating source. But the key there is that for users very often they're getting to our content from sources that are not the library or not library-specific sources. The pie chart is equivalent to this particular bar chart on the graph, so these six bars are six different institutions, University of Denver is one of them. And the bold content at the top, or the bold sections at the top, are the library-originated referrals and you can see in the green, the blue, and the red at the bottom, the stuff that's coming elsewhere. Most of these referrals at these six institutions are coming again from outside the library. They are not coming from library discovery services or the library catalog.

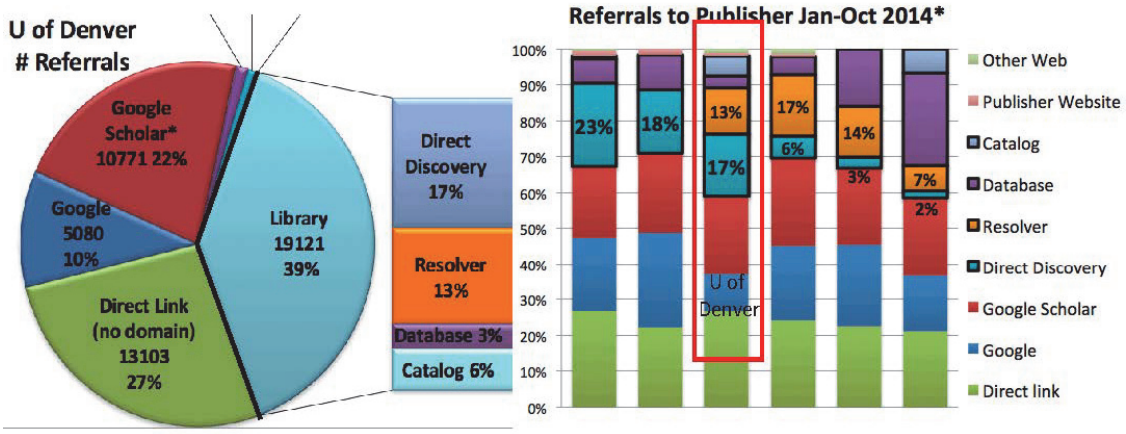

Figure 1. Single publisher referring Site URL data.

https://doi.org/10.5703/1288284316494

And people are getting to content in a lot of different ways. One of these ways is ResearchGate. ResearchGate, as most of us know, is a sort of a social research tool where people can post content, people can share content, and people can ask for contact. There is metadata about this particular article in ResearchGate, but there is also an icon where a user can request that full text. I'm a member of ResearchGate. Many of us are, and one of the sort of annoying features of ResearchGate is that you get a lot of e-mail from them asking you to post stuff, right? So I've got a bunch of notifications here from people who want me to post something. ResearchGate doesn't actually tell, it doesn't help you determine whether you have the rights, as an author, to post a particular article, and very often the things that get posted on ResearchGate are not versions of the article that should be made freely available. They are rogue open access.

And then there is Sci-Hub. Sci-Hub is a tool that is out there with articles that are pirated from all sorts of different sources. Sci-Hub has been quite in the news, including this really detailed study of usage and the history of it that that was in *Science* magazine last year. One of the things that was really interesting about this study is the number of people coming from places where they'd have legitimate access. So, from institutional sites, they're going to Sci-Hub even though they are at universities that have access to a lot of this content. And one of the things that is interesting is that they tell us they're going there for convenience. So, the orange bar on the slide (see Figure 2), the 23% and the 17% there, the convenience factor, so a combined 40% of the users there say they come to Sci-Hub even though they may have access, right? So 51% say they come because they don't have access. Seventeen percent say that they use Sci-Hub because it is more convenient than the library or other sources that they have access to. Twenty-three percent say they object to the profits of publishers. That 40% probably has access, but they are choosing to use Sci-Hub anyway, and this is of 11,000 researchers, this survey. Eighty-eight percent of those surveyed said that they don't actually believe that it is wrong to download pirated papers, so that is an issue that we should all be concerned about, right? That they are using Sci-Hub, and they don't care that it's pirated. They're using Sci-Hub even though they probably have access in other ways.

A recent study shows that in 2013 we actually passed the 50% point for open access content on the web. In April of 2013, 50% of the peer-reviewed articles that had been published in 2011 were available in some form of open access, green or gold, on the web. So, we decided to investigate sort of the broad availability: green, gold, rogue, and pirate, pirated meaning on Sci-Hub, of freely available article content. We randomly selected 300 articles that were indexed in Scopus and published in 2015. A hundred of them are from the arts and humanities, and a hundred of them are from the social sciences, and a hundred are from the life sciences, and all of them, again, randomly selected.

We'll be talking about a few definitions sort of as we go through. I want to just be clear what we mean by these things, by these terms. Availability means the presence of full text in a free version. Right? That we found some full text freely available on the web. We didn't have to login in any way. We searched in four different locations. Our search locations were Google Scholar, Google, ResearchGate, and Sci-Hub. Again, two open sources: ResearchGate, which is sort of rogue in that publishers or authors can deposit a version of the article that may not be a true open access version, and then Sci-Hub where content is pirated. We looked at four different access types across these search locations. There is gold open access, which we defined very broadly as any version that we could get to a free version on the publisher's website. Green open access: We looked in institutional and subject repositories, as well as on author websites, discoverable through Google or Google Scholar. A rogue version is anything that we found on ResearchGate. We did not try to go into ResearchGate and determine which things were legitimate open access versus rogue, so we're just saying if it is on ResearchGate, it is rogue. Pirated means anything on Sci-Hub. Again, on Sci-Hub, some of it is actually open access content. Some of it is content that should not be available.

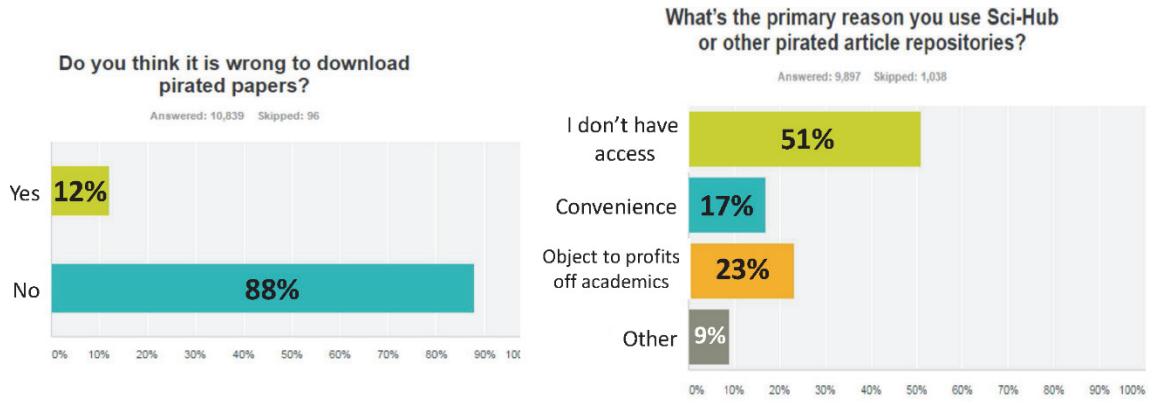

Figure 2. A Science survey of 11,000 researchers. http://www.sciencemag.org/news/2016/05/survey-most-give-thumbs-pirated-papers

We searched each article by title in Google Scholar and in Google. We just did a title search. We didn't do anything further than the title search. We counted the access types. We counted in Google and in Google Scholar whether it was available in gold or green or rogue. In many cases, Google Scholar turns up ResearchGate or Academia.edu results. We counted the number of title match results in each. We counted the number of results with available full text, so how many things could we find full text for when we were not on our campuses using our licensed content? We then searched each article title again in ResearchGate because sometimes ResearchGate turns up in Google Scholar. Sometimes it doesn't, so we searched directly in ResearchGate as well. We searched in Sci-Hub. And then we measured the title match versus the freely available full text results. So, we gathered a bunch of data, and now John is going to come up and talk about some of our results.

John McDonald: Thanks, Michael. This is the best part of the presentation, so, I'm the lucky guy that gets to give you guys all the results. For access type, again, Michael told you access type or, in other words, the source of the full text article, whether it was green, gold, rogue, or pirated, was our first set of results. As far as gold, green, and rogue, we had just a few simple research questions. Basically, how many are gold out of our article sample? How many are green, and then where are they green? Are they green in institutional repositories, subject repositories, or on author websites? And then how many are in the rogue and the pirate systems? For rogue systems, we did ResearchGate and Academia.edu, and for pirated, it was Sci-Hub. And a note about Academia.edu: You can't search it directly, so we only got results through Google results, so you'll see the results in one of the next slides.

So, here is the verdict. Out of our sample articles available in gold OA, we found that a total of 80 out of our 300 articles were available gold OA on the publisher's website. That's 26% of the sample, and across the disciplines, it ranged from a nonsurprising 23% in Arts and Humanities up to 32% in the Life Sciences (see Figure 3).

Articles available via Gold OA	
Discipline	Publisher Websites
Arts & Humanities	23
Social Sciences	25
Life Sciences	32
Total	80/300 (26%)

Figure 3. Articles available via Gold OA.

Then for green OA, the articles available green OA overall, we found that institutional repository green OA accounted for 9% of the articles were found in institutional repositories. That was relatively surprising to us that institutional repository copies

were not as discoverable as we expected. Subject repositories were a little bit better but still not great at 14% overall, and not surprisingly probably to all of the librarians in the room, the author websites self-archived were not very discoverable at all. We only found 10 articles out of our sample in total (see Figure 4).

As far as our rogue systems, ResearchGate and Academia.edu, we found that 30% of the total sample was available via ResearchGate, and the Arts and Humanities are not very accessible in ResearchGate as open access versions, but the Social Sciences ended up with 36% and Life Sciences 44%, so probably what everybody would expect. As far as Academia.edu, again, I didn't put a percentage on the table here because we weren't accessing Academia.edu directly, so there could be additional items in there that are open access, but this is what we got from our Google and Google Scholar results. Overall, the total for both of these rogue systems together were 111 articles, so 37% (see Figure 5).

And the grand total for all open access sources ended up being 166 of the 300 articles; we could find at least one version of an open access article. Arts and Humanities was just below 50%, Social Sciences very high at 60%, and then Life Sciences at 57%. And these results match the earlier research results that have been published in the literature that write about 50%, 50 to 60% of recently published literature is available in an open access form (see Figure 6).

Articles Available via Green OA				
Discipline	Institutional Repository	Subject Repository	Author Website (Self-Archived)	Total Articles
Arts & Humanities	6	4	5	13
Social Sciences	14	10	3	19
Life Sciences	7	27	2	27
Total	27 (9%)	41 (14%)	10 (3%)	59 (20%)

Figure 4. Articles available via green OA.

	Articles available in Rogue Systems		
	ResearchGate	academia.edu	Total Rogue
Arts & Humanities	11	20	26
Social Sciences	36	9	40
Life Sciences	44	5	45
ALL	91 (30%)	34	111 (37%)

Figure 5. Articles available in rogue systems.

To contrast that with Sci-Hub, we searched all of the articles in Sci-Hub, and we came up with an astounding 87% of the articles were available in Sci-Hub and equally across all the disciplines. We found 86 of our article in Arts and Humanities were available in Sci-Hub, and 87 in Social Sciences, and 87 in the Life Sciences (see Figure 7).

Looking at this availability then as a bar chart (see Figure 8), on one slide you can see then that gold open access via publisher websites, we ended up with 80 of the articles total. Green open access in all locations was not as available as gold open access, but ResearchGate in the blue bar—ResearchGate and Academia.edu actually performed pretty well with 111 of the articles. Overall, the Arts and Humanities are not well served by ResearchGate and Academia.edu but pretty comparable in gold open access at least. The Life Sciences have higher percentages, as most people would expect, but the Social Sciences performed pretty well, especially in ResearchGate.

Discipline	All OA Sources
Arts & Humanities	49
Social Sciences	60
Life Sciences	57
Total	**166/300 (55.3%)**

Figure 6. Recently published literature available in Open Access form.

Discipline	Pirated Articles available in Sci-Hub
Arts & Humanities	86
Social Sciences	87
Life Sciences	87
Total	**260/300 (87%)**

Figure 7. Pirated articles available in Sci-Hub by discipline.

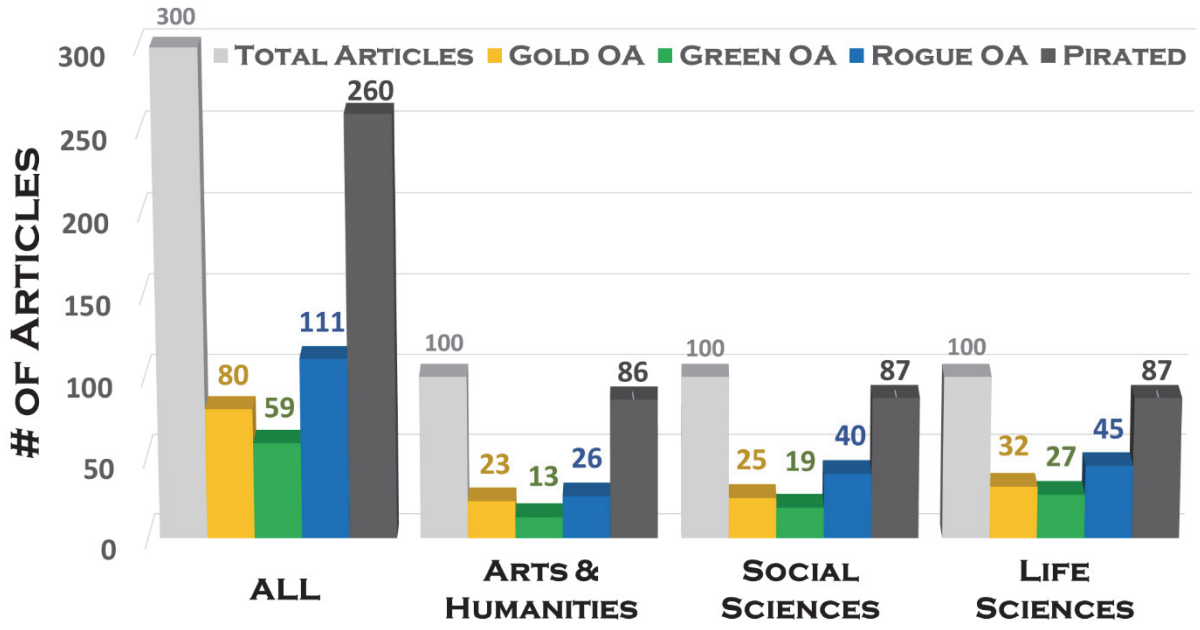

Figure 8. Availability by access type.

Then we added the black bars here to show those compared to Sci-Hub, and you can see that 260 total articles, the 86, 87, and 87 across the three broad disciplines.

Now, those were the total articles. We also wanted to look at the additive availability by the article source so, for example, gold open access we found 80 articles by gold open access to, if as publishers and librarians, we feel like that that is the most legitimate variety of open access that there is, 80 of the articles were available gold open access. And then if you start to look at things that were green open access but not gold open access, so how many additional articles were available in an open access version that weren't available gold, but they were available green? We found that additional 24 articles, so then we are up to 104 out of our 300 article sample.

Moving forward, we looked at what was available in our rogue systems that wasn't otherwise available in gold or green, and we found an additional 59 articles. Then if you go—we found 59 in the rogue system, and then if you add in Sci-Hub to complete your journal article searching, then you found an additional 115. Overall, all versions of freely accessible journals ended up over 90%, so our users could relatively easily discover about 90% of the articles in our sample. And you will see that even though Arts and Humanities is not as well represented in gold, green, and the rogue systems, Sci-Hub makes up for it with great coverage of the Arts and Humanities as well. Hey, if you're going to steal articles, you might as well do it from the Arts and Humanities journals, too, right?

We also wanted to look at, as Michael told you earlier, we were looking at search location. So, generally looking at how users, scholars, mostly faculty and students, are actually finding this content. We wanted to look at Google Scholar, Google, ResearchGate, and Sci-Hub as the search location for all of the articles. And a little note about methodology, we did start off with Google Scholar, making a broad assumption that most academics know Google Scholar and may start with Google Scholar. Some institutions even use Google Scholar as their discovery system. We started with Google Scholar, and we were looking at search results for our articles, and we looked at the "All Versions" button below every article. They collate all the versions that they think they found, that Google Scholar thinks are the same article, and they put them together. So we found the results, and then we expanded to look at all 10 versions, and we also noted the PDF view. Google Scholar is promoting access to freely available articles and legitimate open access by directly linking to PDFs that they can find. You will find that on the right-hand side of

search results. And we found that the "All Versions" for most articles out of Google Scholar, the overall average was 3.74. So Google Scholar is finding three to four versions of every single article. Unsurprisingly, the Arts and Humanities is not as well represented with only 2.5, and the Life Sciences much better represented with five. And then we found that Google Scholar will provide you access as a search location to over 40% of the journal articles in our sample, so you can get to it open access from Google Scholar for 122 of our articles.

And then when we then progressed looking at doing the same searches in Google, we found that it was the exact same number of articles that you can find through Google, 122 of our 300, and they were not always the same articles. So, the 122 we found in Google Scholar were a different set than you could find in Google. So, that's why users should actually search through both of them. Fewer number of title matches in Google; they don't collate the matches, but when you do article level searching here, you will see multiple versions come up and in the Arts and Humanities. It was just below 3, Social Sciences right at 3, and life sciences at above 3.5.

Looking at these results, availability by search location in one chart (see Figure 9), again Google Scholar is the blue bars, and we found 122 of our articles overall, Google with also 122, ResearchGate was 91 articles we found, and we put Sci-Hub on here also to underscore the total volume that you can get through Sci-Hub is 260. Google Scholar and Google operate almost equally in the discoverability of this content, and ResearchGate functions really well for the Social Sciences. You will see there were 36 articles found through ResearchGate in the Social Sciences as compared to 39 in Google Scholar and 40 in Google, and in the Life Sciences, it is even closer. ResearchGate does not have great coverage in content in the Arts and Humanities right now.

Looking at the additive availability by search location again, and we've got two different versions we're going to go through here. Google Scholar provided access to 122 of the articles. If you then move on to Google and limit out the ones you've already found, you find an additional 32 through Google, so 154 now of our sample, so just over 50% of the content was available by just searching Google Scholar and Google. Then if you move on to ResearchGate, you'll find eight additional articles that you didn't discover before, and then again Sci-Hub, you will end up finding basically the rest of the sample. So, up to over 90% of the total articles were available if you use all four of these methods.

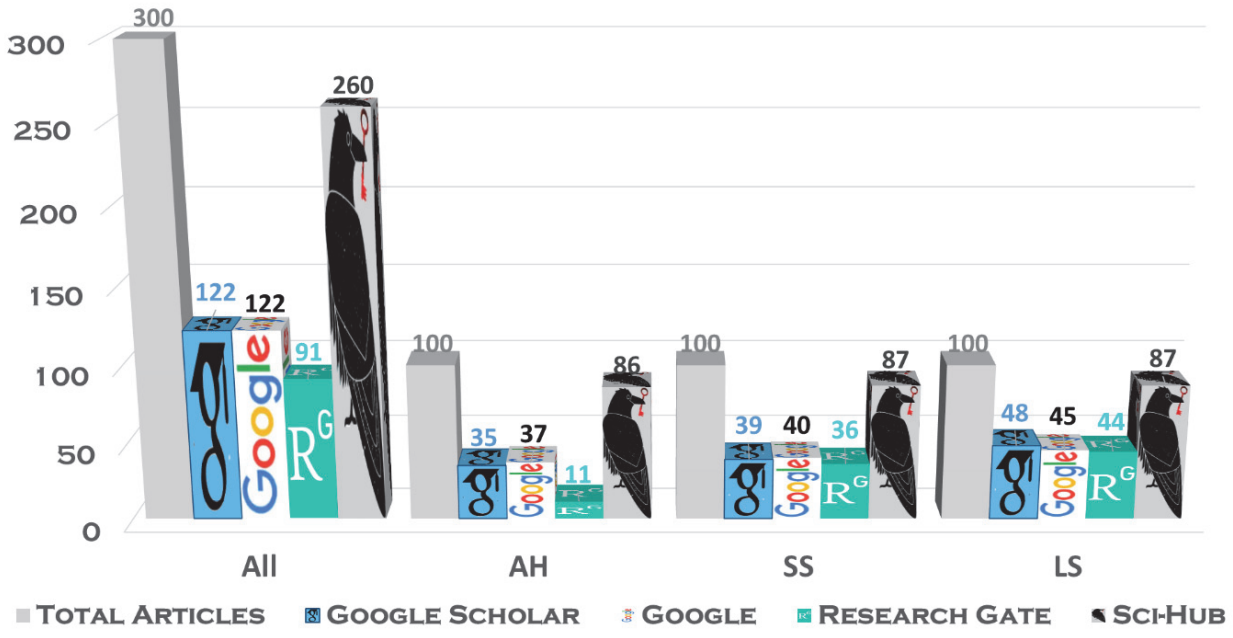

Figure 9. Availability by search location.

Moving backwards, though, we wanted to look at it if authors, faculty that are on ResearchGate and use ResearchGate as a discovery mechanism, if they actually started from ResearchGate, what is the additive availability by using ResearchGate first, and then moving on to others. You would find 91 of the articles via ResearchGate, and then when you go to Google Scholar, you find an additional 54, an additional 18 from Google, and then basically the rest from Sci-Hub. I'll turn it over to Jason to conclude.

Jason Price: John says his was the best part, but I think this is the best part, although it is also a sensitive subject in some ways. I want to be a little provocative and lead into some discussion, so we're looking forward to that.

So, in conclusion, it is hard to follow the rules. If you stick with the open access versions of articles, you are limited to somewhere in between 20 and 40% depending on the source of that version, and in fact, I guess 20 to 25% on the classic rights-appropriate open access. If you go into the rogue open access, which is potentially much less rights-appropriate, but you increase the number, but if you want to go just one place and get the most possible freely available articles, as a researcher who doesn't think that it is wrong to download pirated articles, you're going to go to Sci-Hub. Starting with Google Scholar and supplementing that by Google is slightly a better strategy than starting with ResearchGate, but you can kind of move both ways, and even though you see 30, 20, 40%, you can get up to a higher number if you use one and then progress on to the next, if need be. Again, starting with Sci-Hub and bypassing the legitimate search options entirely gives the quickest and best results. And an obvious conclusion from that is that libraries and publishers should be concerned if our users decide to go here instead of using the contents we are licensing, that is a huge problem, one that we need to recognize and not ignore.

Before I go on to some of the potential applications of this, I want to talk about one next step that we haven't taken that we think is really important, and that is to examine both OA discoverability and availability in library discovery systems. So, this graph (see Figure 10) looks at the four most popular discovery systems, and the blue bars are articles that aren't available in open access, and the gray bars are those that are. So, the question is if you just drop that title in that discovery system, is it going to be indexed? This is not necessarily a test of OA versus non-OA, but that's the intent: Is OA content less well indexed in library discovery systems? That is a relatively important question. We didn't see strong trends toward that, but we did have this discoverability side, although there could be something underlying this. More importantly, potentially, is how effective are library linking tools at providing the full text access to open access articles? So, if you find it in your discovery system but you don't have licensed access to it, how commonly do our systems lead to that full text? We expect the answer to be not nearly as commonly as they are actually available out there on the web, but we would like to actually design a study to look at that in a little more detail. And I think we, I work for a library consortium of very small libraries, many of whom do not have site license access to a lot of this content, and I think doing this work with them, examining that some more, might open up some possibilities.

The theme of the conference: "Roll With the Times or the Times Will Roll Over You." This theme and our presentation I think really fit well this year. The times, led by faculty who are sharing articles in ways which may or may not be rights-appropriate and who feel like it is fine to download pirated papers and are going there, that's the times are pushing forward, and we need to not ignore these things. We need to recognize them and think about how we can react and respond appropriately.

I have three puns for you, and I'm going to give an example of each. The first is "Collar Google Scholar?" The second is "Emulate ResearchGate?" And the third thing to do in response this is: "Don't ignore that there is a Sci-Hub Pirate Club out there."

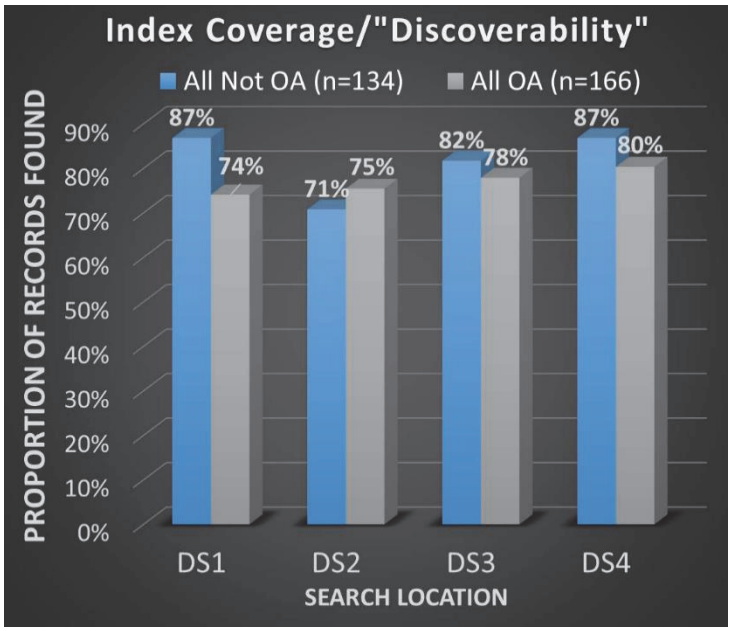

Figure 10. Index coverage/"discoverability."

"Collar Google Scholar," what do I mean? Maybe we should be linking to Google Scholar results from our open URL resolvers in order to leverage more open access full text. That is a possibility and/or drawing Scholar open access text links into the results menu when they are available. Google Scholar is actually doing this. They have created a plug-in which allows you to highlight text, hit a button, and then pull up this sub-window on the right-hand side, and that green button is an open access button. If you are a researcher and you've added this plug-in, which they are now advertising underneath their search results, they're making it obvious to faculty that this exists, and they are leveraging and making these open access links much more visible and likely to be used by researchers. They are even going to the point where on a publisher page where it says, "Purchase this Article," they are pointing out that potentially, even without selecting any text, if you hit that Google Scholar extension button, it shows you the open access version of that article instead of the one that you might pay for as a researcher. So, we need to recognize that Google Scholar is leveraging these links, and we need to find ways to leverage these links. I just learned today, actually, that Elsevier has created an article-level knowledge base that indicates which of the articles are freely available and which are not. So, think hybrid journals: You can't use the title and figure out whether it is open access or not because there's both kinds in there. They have an API which is freely available to folks to potentially, if you have a DOI, you check it and it will say, "Yes, this is open access," or "No, it's not." We could put an article level link in our results pages for folks who don't subscribe to those journals but can get access to them. These are the kinds of things that I think we need to be doing to be keeping up with the times.

Second example: "Emulate ResearchGate." So, this is something a library is already doing: Include metadata for all faculty publications in institutional repositories, even if the OA copy is not available and even potentially if it never will be, and allow users to request a copy through the institutional repository listing. So, the text on this is small, but you'll get the idea. This is the University of Liege (see Figure 11). This is their institutional repository. They just have an abstract. The bottom of the page on the left-hand side shows that there is restricted access to this article, but there is a PDF in there. On the right-hand side, they have a button to request a copy. Does that sound familiar? That is what you do in ResearchGate. Here is a library doing that. When you hit that button, it tells you if you are from the University login. If you are not, here are the rules, but you can ask the author of this article for a copy. That's what they are doing with their institutional repository. I think this is emulating the fact that ResearchGate actually covers—it has listed—nearly 100% of the articles we looked at. I think that is important otherwise our institutional repositories really don't cover an extensive portion of our faculty publications.

The third example is less of an example and just something that when we found this as part of doing this research, I was floored. Here is a big thing that is going on that I think we should know about and recognize is happening. Remember 88% of researchers did not think it is wrong to download pirated papers, and 87% of the papers are pirated and available through Sci-Hub. 87%, right? There is a plug-in that if you go to Sci-Hub's site, if you look for an article and it's not found, it gives you a link to install this Google Chrome extension. Now, that said, it is a developer mode. It's a little funky kind of thing, but because Google has not endorsed supporting Sci-Hub, they are not adding this extension into their publicly available content. What you'll notice if you look closely is that down, you probably can't see the URL, but it points to the article from a Google Scholar interface in Sci-Hub. When you click that title there in a Google Scholar interface, you go directly to a Sci-Hub pirated version of that article instead of going to where you normally. If you had the Google Scholar without the plug-in, you would go to your campus's licensed access if you're on campus. This makes it extremely convenient to access 87% of the articles published in 2015 across the disciplines. That is scary to me but also something that I think we can't ignore and need to address. So, with that, I'll open it up for questions, comments, thoughts.

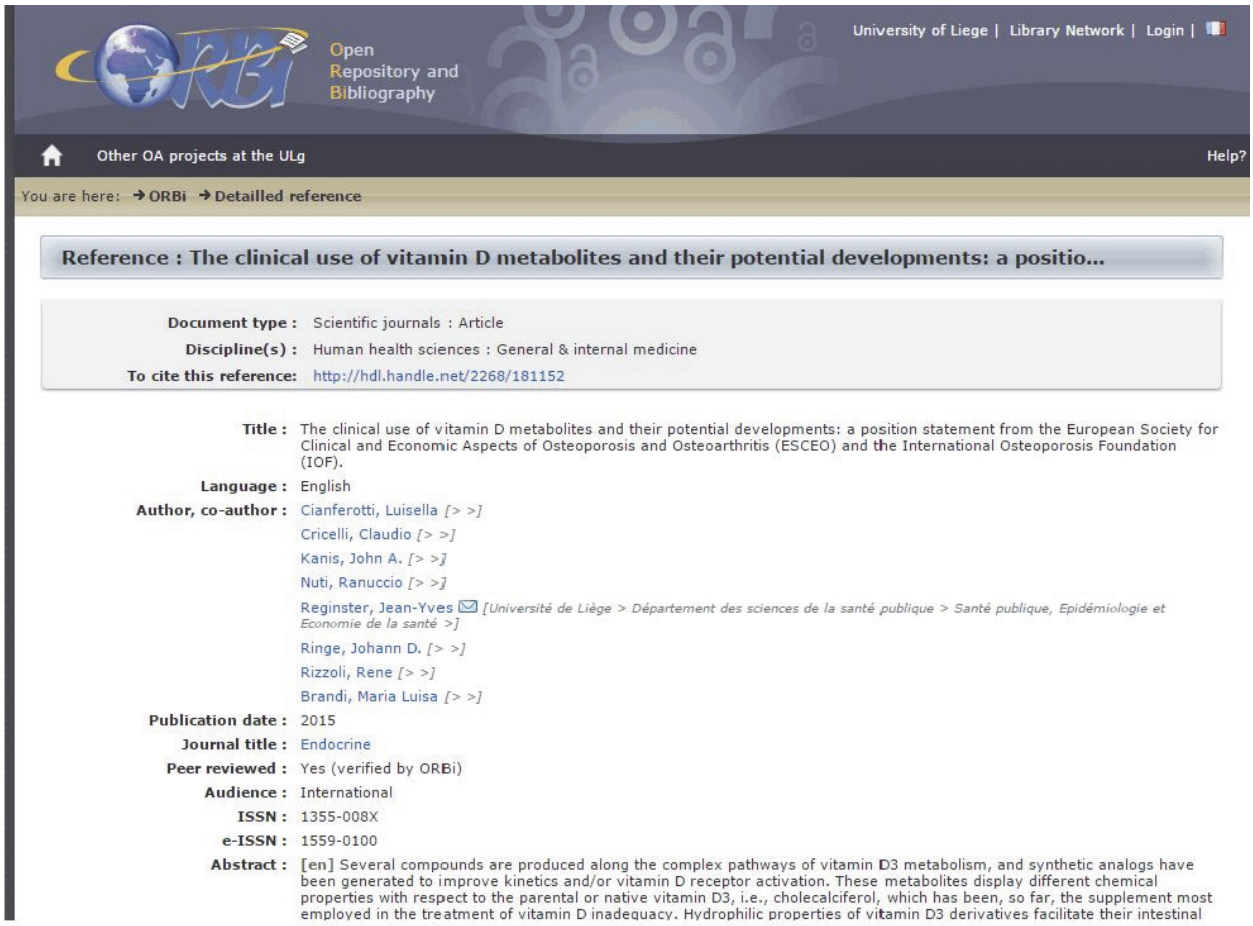

Figure 11. University of Liege institutional repository.